JCCES®

JOUVE-CEREBRALS CRYSTALLIZED EDUCATIONAL SCALE

Technical and Interpretive Manual

Cognitive Acumen Index (CAI) · Verbal Acumen Index (VAI)

DOCUMENT VERSION	2026.2
PUBLISHED	May 26, 2005
LAST UPDATED	July 2026
TEST VERSION	JCCES v2025.1
TEST AUTHORS	Jouve, X., Wai, J., Watterson, J., Woodley, M., Faverio, M. C., Lindgren, K., Whitley, J., & Freitas, N. (2005–2010)
MANUAL AUTHOR	Xavier Jouve, Head of Research
CONTRIBUTING EDITOR	Priya Sharma, Head of Assessment Development

CONTENTS

1 The Instrument and Its Measurement Philosophy
2 The Published Evidence Base
3 Scale Architecture and Score System
4 The Item Pool: Formats, Keys, and Calibration
5 Administering the Scale
6 From Responses to Indices: Scoring and Norming
7 Reliability and Precision
8 Internal Structure
9 Criterion Evidence
10 Score Interpretation and Profile Analysis
11 Protocol Review
12 Uses, Populations, and Boundaries of the Evidence
— References

CHAPTER 1

The Instrument and Its Measurement Philosophy

The Jouve-Cerebrals Crystallized Educational Scale (JCCES; Jouve et al., 2023) is a 129-item, untimed, open-ended measure of crystallized intelligence for online self-administration. Its three subtests, Verbal Analogies (VA), Math Problems (MP), and General Knowledge (GK), sample the acquired-knowledge domain from three directions, and two composite indices, the Cognitive Acumen Index (CAI) and the Verbal Acumen Index (VAI), summarize performance on the conventional deviation-IQ metric. The instrument grew out of the Cerebrals Cognitive Ability Tests, built in 2004 by a multidisciplinary team of the Cerebrals Society, and reached its consolidated present form in the revision documented by Jouve et al. (2023).

What the scale intends to measure is the construct the Cattell–Horn–Carroll taxonomy calls Gc: the store of lexical knowledge, general information, and schooled reasoning that a person accumulates through education and cultural experience (Carroll, 1993; Schneider & McGrew, 2018). Crystallized ability holds a privileged place in applied assessment: it is among the most replicated dimensions in the factor-analytic literature (Horn & Blankson, 2005) and among the strongest single predictors of academic attainment (Floyd et al., 2003).

Three commitments distinguish the JCCES from recognition-format measures of the same domain, and each is a measurement decision rather than a stylistic one. First, every answer is typed, never selected, so chance success is structurally excluded and item difficulty can extend to extremes that a multiple-choice floor would contaminate. Second, no clock runs: the score reflects what the examinee knows and can work out, not retrieval speed, which keeps the operational construct aligned with the theoretical one (Horn & Cattell, 1966). Third, when a question genuinely admits several defensible answers, the key credits them all, and which answers count as defensible is decided from response data under a documented procedure, not by editorial fiat (chapter 4).

This manual is the technical and interpretive documentation for the operational first edition. Its organization follows the sequence the *Standards for Educational and Psychological Testing* recommend (AERA, APA, & NCME, 2014): the instrument and its research base (chapters 1–2), the scale and its development (chapters 3–4), administration and scoring (chapters 5–6), the psychometric evidence (chapters 7–9), and interpretation, protocol review, and boundaries of use (chapters 10–12).

CHAPTER 2

The Published Evidence Base

This edition of the manual adopts a strict documentation policy: it presents no analysis of its own. Every coefficient, sample size, and model statistic in the chapters that follow is quoted from a published,

peer-reviewed study of the scale and cited where it is used, so that any number can be traced to its source and read in context. The manual's job is to organize that record for the test user; the record itself lives in the literature.

Table 2.1 maps the evidence base. Two large samples anchor it: the original calibration corpus of 1,079 complete protocols (Jouve, 2023) and the standardization corpus of 1,553 complete protocols on which the current norms, the factor structure, the discrepancy base rates, and the scoring methodology were established. Around these sit five external-criterion studies spanning clinical, educational, and military score traditions, a multidimensional scaling study, an item-level study of the pooled JCCES and What's Next banks, and a pilot calibration of the forthcoming second edition.

Table 2.1

Map of the published evidence base of the JCCES

Evidence	Study	Sample	Establishes
Instrument and calibration			
Canonical reference	Jouve et al. (2023)	—	The instrument, its revision, and its authorship
Calibration and reliability	Jouve (2023)	*N* = 1,079	2PL calibration, item statistics, internal consistency
Answer-key construction	Jouve & Sharma (2026a)	*N* = 1,553	Key coverage, accuracy, normalization policy
Option-curve scoring	Jouve & Sharma (2026c)	*N* = 1,553	Answer multiplicity, dichotomous-scoring evidence
Parameter cross-validation	Jouve & Sharma (2026b)	*N* = 1,553	Independent recovery of the item parameters
Composite conversion	Jouve (2026a)	*N* = 1,553	Conversion family, archived tables, index precision
Structure and interpretation			
Factor structure	Jouve (2026b)	*N* = 1,553	Three-factor form, dominant general dimension

Evidence	Study	Sample	Establishes
Subtest scaling	Jouve (2016)	*n* = 57	Content and instrument geometry against the ACT
Pooled item bank	Jouve (2025)	*n* = 47	Item-level *g*-saturation with an external test
Discrepancy base rates	Jouve (2026h)	*N* = 1,553	Normative rarity of profile differences
External criteria			
Clinical verbal indices	Jouve (2026d)	*N* = 119 / 56 / 16	Convergence with RIAS, WAIS-IV, WISC-III
SAT	Jouve (2026f)	*N* = 105	Convergence with a national admissions battery
GRE	Jouve (2026c)	*N* = 65	Multitrait–multimethod convergent/discriminant contrast
AFQT	Jouve (2026e)	*N* = 62	Convergence and mean-level agreement, military composite
GAMA	Jouve (2026g)	*N* = 57	Discriminant separation from a fluid-figural battery
Development			
Second edition (pilot)	Jouve (2024)	*N* = 62	JCCES-II pilot calibration and provisional equating

Note. Samples are complete-protocol counts as each study reports them. 2PL = two-parameter logistic; RIAS = Reynolds Intellectual Assessment Scales; WAIS-IV = Wechsler Adult Intelligence Scale–Fourth Edition; WISC-III = Wechsler Intelligence Scale for Children–Third Edition; AFQT = Armed Forces Qualification Test; GAMA = General Ability Measure for Adults; ACT and SAT and GRE are the standardized educational tests of those names.

Reading the manual against Table 2.1 shows the division of labor. Chapters 4 and 6 rest on the calibration and scoring-methodology studies; chapter 7 on the two large-sample reliability records; chapter 8 on the structural studies; chapter 9 on the five criterion studies; chapter 10 on the base-rate study; chapter 12 on the pilot of the second edition. Where a chapter states a procedural rule rather than an empirical

finding, the rule is operational policy of the testing platform and is labeled as such.

CHAPTER 3

Scale Architecture and Score System

Verbal Analogies presents 41 items in relational form: a pair of words fixes a relation, and the examinee supplies the word that completes a second pair under the same relation. Because a relation can often be completed by several near-synonyms, VA items are the scale's most multiple-answer format. Math Problems presents 32 quantitative problems, from arithmetic through elementary algebra and geometry, each with a single correct numerical answer that must be produced, not picked from options. General Knowledge presents 56 direct factual questions spanning history, science, geography, and the arts; many admit more than one correct answer, and the key credits the documented alternatives (Jouve et al., 2023; Jouve, 2023).

Seen through the CHC lens, the three formats deliberately triangulate Gc: VA reaches lexical knowledge through reasoning, GK reaches the declarative store directly, and MP reaches quantitative knowledge in use. That MP would also lean toward reasoning ability is expected under the investment account of crystallized intelligence, and the structural evidence of chapter 8 confirms both the shared core and MP's distinct quantitative footprint (Carroll, 1993; Schneider & McGrew, 2018).

Table 3.1 lays out the score system, and Figure 3.1 traces how a protocol moves through it. Raw subtest scores are converted to standard scores against the standardization sample; the standard scores feed the two indices. The CAI, built from all three subtests, is the scale's primary interpretive score; the VAI, built from VA and GK, serves referral questions confined to the verbal-crystallized core. Both indices report on the deviation-IQ metric, so a JCCES result reads in the frame of reference shared by the Wechsler-family instruments (Wechsler, 2008), with the essential caveat that shared metric is not shared norms: cross-

instrument agreement is demonstrated empirically in chapter 9, not assumed from the scale's appearance.

Table 3.1

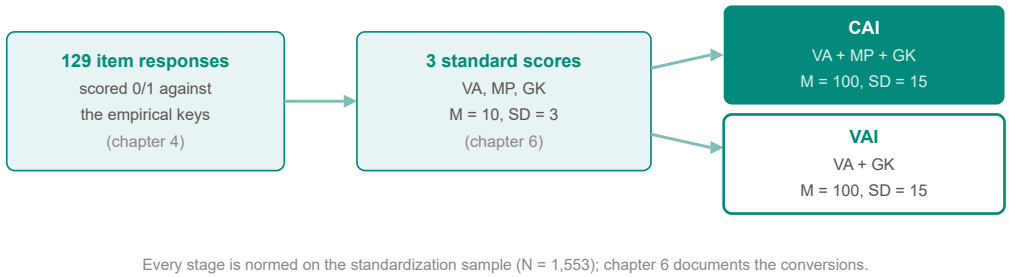
The JCCES score system

Score	Built from	Metric	Reliability	Role
VA standard score	41 items	*M* = 10, *SD* = 3	.853	Verbal-reasoning facet
MP standard score	32 items	*M* = 10, *SD* = 3	.914	Quantitative facet
GK standard score	56 items	*M* = 10, *SD* = 3	.942	Knowledge-breadth facet
Cognitive Acumen Index	VA + MP + GK	*M* = 100, *SD* = 15	.963	Primary interpretive score
Verbal Acumen Index	VA + GK	*M* = 100, *SD* = 15	.953	Verbal-crystallized core

Note. All scores are normed on the standardization sample (N = 1,553). Subtest coefficients are α on the dichotomous items (Jouve, 2026b); index coefficients are stratified α (Cronbach, Schönemann, & McKie, 1965; Jouve, 2026a). Chapter 7 discusses the reliability evidence; chapter 6 documents the conversions. VA = Verbal Analogies; MP = Math Problems; GK = General Knowledge.

Figure 3.1

From item responses to reported indices


129 item responses
scored 0/1 against
the empirical keys
(chapter 4)
3 standard scores
VA, MP, GK
M = 10, SD = 3
(chapter 6)
CAI
VA + MP + GK
M = 100, SD = 15
VAI
VA + GK
M = 100, SD = 15
Every stage is normed on the standardization sample (N = 1,553); chapter 6 documents the conversions.

Note. VA = Verbal Analogies; MP = Math Problems; GK = General Knowledge; CAI = Cognitive Acumen Index; VAI = Verbal Acumen Index.

CHAPTER 4

The Item Pool: Formats, Keys, and Calibration

Three published studies of the standardization corpus, together with the original calibration study, document how the item pool works: what the items ask, how their keys are built and audited, and how their psychometric parameters were estimated and verified (Table 2.1). This chapter assembles that record.

What the Items Ask

One published example from each subtest fixes the formats:

Hot : Cold :: Fast : _____ — accepted answers include Slow*,* Sluggish*,* Leisurely.

If 3x + 7 = 22, what is x? — answer: 5.

Name a planet in our solar system with rings. — accepted answers: Saturn, Jupiter, Uranus, Neptune.

How many answers an item truly has is an empirical property of the domain, and it differs sharply by format: analyzed over the full standardization corpus, VA items accept a mean of 5.8 semantically distinct correct answers, GK items 3.6, and MP items essentially one, with 31 of the 32 mathematical items admitting exactly one (Jouve & Sharma, 2026c). The multiple-answer keys are therefore not a leniency but a description of the construct's semantics.

How the Keys Are Built and Audited

Operational keys are constructed by frequency: candidate response forms are ranked by how often examinees produce them, and forms are admitted by a documented decision rule. The published construction study shows why so compact a key suffices: correct responses concentrate heavily on canonical forms, the single most frequent form covering 98.5% of correct responses on mathematical items, 76.3% on

knowledge items, and 64.2% on verbal-analogy items. Keys averaging 1.4, 4.2, and 7.5 entries per item for the three formats reproduce operational scoring on held-out examinees at 99.97% accuracy, 100.0% precision, and 99.86% recall (Jouve & Sharma, 2026a). Table 4.1 collects these properties beside the calibration statistics.

Two further findings from the same research define operational policy. Matching tolerates case and whitespace variation but deliberately preserves punctuation and internal spacing, because stripping them was shown to collapse distinct answers together and cost precision, down to 98.87% on mathematical and 98.79% on knowledge items. And the key construction doubles as an audit instrument: 2.36% of nominally incorrect responses in the corpus proved to be exact surface variants of credited answers, concentrated in GK (4.94%), and the audit localizes each case for maintenance; re-scoring with the audited key moves the external validity coefficients of chapter 9 by at most .012 (Jouve & Sharma, 2026a; Jouve & Sharma, 2026c).

Whether responses should be scored beyond right-versus-wrong has also been answered empirically. Kernel-estimated characteristic curves for every distinct response form, the same technique that first justified crediting popular alternative answers in the calibration study, where roughly 20 to 30% of examinees across the ability range chose a second valid answer (Jouve, 2023), show that nominal scoring of selected items raises MP internal consistency from $\alpha = .911$ to .931 yet lowers the corresponding criterion correlations (SAT Mathematics falls from .777 to .749), and that grading among correct answers helps nowhere. The deployed dichotomous rule rests on that result (Jouve & Sharma, 2026c).

Calibration and Its Independent Verification

Item parameters were estimated under a two-parameter logistic model on the 1,079-protocol calibration corpus, with convergence for every item (Jouve, 2023). The pool's defining property is breadth: within each subtest, proportion-correct difficulty runs from near-floor to near-ceiling (Table 4.1), so a single fixed form carries measurement across the ability continuum, and the interleaved item order distributes that

breadth over the whole session rather than presenting an ascending staircase (Jouve, 2023). Item–total discrimination is positive for every item. The classical account of such scores is given by Lord and Novick (1968); the IRT machinery follows Embretson and Reise (2000).

Table 4.1

Properties of the item pool and its answer keys, by subtest

Property	VA	MP	GK
Calibration (*N* = 1,079; Jouve, 2023)			
Items	41	32	56
Difficulty range (*p*)	.003–.983	.002–.844	.025–.890
Polyserial r_{it} range	.247–.693	.437–.683	.361–.780
Answer keys (*N* = 1,553; Jouve & Sharma, 2026a, 2026c)			
Distinct valid answers per item, mean	5.8	1.06	3.6
Key entries per item, mean	7.5	1.4	4.2
Held-out accuracy	99.96%	99.97%	99.96%
Held-out precision	100.0%	100.0%	100.0%

Note. Ranges are minima and maxima over the items of each subtest; p = proportion answering correctly; r_{it} = polyserial item–total correlation. Key accuracy is measured on a held-out 30% of examinees against the operational reference scores. VA = Verbal Analogies; MP = Math Problems; GK = General Knowledge.

The calibration itself has been verified by a method with no machinery in common with likelihood estimation. A group-theoretic estimator that reads item parameters from score-group invariants of the response matrix, applied to the 1,553-protocol corpus, reproduced the marginal-maximum-likelihood parameters at $r = .9994$ for difficulty (RMSE = 0.086 logits) and $r = .98$ for discrimination, and matched the reference model's prediction of held-out examinees within 0.02 nats (Jouve & Sharma, 2026b). When two unrelated estimation principles agree to that degree, the parameters belong to the data, not to the software.

CHAPTER 5

Administering the Scale

Administration is online, unproctored, and self-paced, in one session wherever practicable. The examinee needs a quiet setting free of interruption, a device with a physical or on-screen keyboard, scratch paper for the mathematical items, and working proficiency in written English; typical completion time is 60 to 90 minutes, though nothing ends the session at any duration. Conditions of testing belong to the evidential record of a score, and the general obligations for standardized administration apply (AERA, APA, & NCME, 2014; Jouve et al., 2023).

What Examinees Are Told

The instructions communicate the scale's ground rules before the first item: there is no time limit; items arrive in mixed rather than ascending difficulty; wrong answers cost nothing beyond the missed credit, so an attempt is always better than a blank; and answers are entered as plain words or numbers, with the matching rules forgiving capitalization and stray spacing (chapter 4), so no effort should go into formatting. For Math Problems, only the final answer is entered; the working stays on scratch paper.

The Session as Behavioral Record

With no examiner present, the session log stands in for chair-side observation, and it is read at three levels: setting (device, connectivity, interruptions), pacing (continuous versus paused work, latencies by item), and compliance (navigation, completion, care of entry). The reviewer's interest is procedural throughout; inferring motive or diagnosis from a log is outside its evidential reach (AERA, APA, & NCME, 2014).

Certain patterns inform the protocol review of chapter 11 without deciding it. A block of blanks inside one subtest can mean

disengagement or content avoidance, and takes its meaning from the validity indicators. Runs of implausibly fast answers raise the rapid-responding question below. Heavy answer revision usually signals deliberation rather than trouble. Recurring entry errors of a typographical kind point at the keyboard, not the construct, and prompt a check that normalization behaved as designed.

Rapid Responding

Where per-item timing exists, responses faster than a minimal read-and-solve threshold are flagged in the tradition of response-time effort analysis (Wise & Kong, 2005), the pattern-level screens of chapter 11 are recomputed without the flagged items, and the report discloses the count removed. Without timing, the pattern-based screens do the whole job.

Stating the Conditions in a Report

Every written report opens its results section with a conditions statement of this form:

> "Testing was completed online in one continuous sitting; the session log shows no interruptions, connectivity failures, or unusual entry behavior, and response pacing was unremarkable. Conditions are judged adequate for interpretation (AERA/APA/NCME, 2014)."

CHAPTER 6

From Responses to Indices: Scoring and Norming

Scoring proceeds in the three stages of Figure 3.1. Responses are matched dichotomously against the keys of chapter 4; raw subtest totals become standard scores ($M = 10$, $SD = 3$) through conversions normed on the standardization sample; and the standard scores combine into the CAI and VAI. The composite stage, its rationale, and its published

precision are documented in the conversion study (Jouve, 2026a), which archives the adopted conversions; the account below follows that source.

The Composite Conversion

A composite begins as the standardized sum of its subtest standard scores. With n subtests $S_1, \ldots, S_n$ and pairwise intercorrelations summing to $R = \sum_{i<j} r_{ij}$,

$$z \;=\; \frac{\sum_{i=1}^{n} S_i \;-\; 10\,n}{3\,\sqrt{n + 2R}}\,,$$

the classical composite standardization of Tellegen and Briggs (1967). The standardized sum then reaches the reporting metric through the generic monotone exponential conversion, GMEC-1:

$$G(z) \;=\; \int_0^{z} \exp\!\left(2At + Ct^{2}\right)\,dt, \qquad Q \;=\; 100 + 15\,G(z).$$

Note. The curvature coefficients are fixed functions of the number of subtests and their mean intercorrelation, estimated once from published battery norms and never refitted per instrument (Jouve, 2026a). At the center the conversion is the identity—$G(0) = 0$, $G'(0) = 1$—so curvature enters only away from the mean. Reported indices are rounded to integers.

The inputs are properties of the standardization sample: subtest intercorrelations of .599 (VA–MP), .774 (VA–GK), and .566 (MP–GK) at N = 1,553 (Jouve, 2026b), which give the per-index structural constants of Table 6.1.

Table 6.1

Structural inputs of the two composite conversions

Index	*n*	*R*	$\bar{\rho}$	Attainable sums
CAI (VA + MP + GK)	3	1.939	.646	19–57

Index	*n*	*R*	$\bar{\rho}$	Attainable sums
VAI (VA + GK)	2	0.774	.774	13–38

Note. From Jouve (2026a), on the standardization sample. n = number of subtests; R = sum of pairwise intercorrelations; $\bar{\rho}$ = their mean. The sums span the documented subtest standard-score ranges (VA and MP 6–19, GK 7–19). CAI = Cognitive Acumen Index; VAI = Verbal Acumen Index.

Why Not a Linear Map, and Why Not a Percentile Table

Two simpler routes are conventional, and the conversion study weighs both (Jouve, 2026a). The first is the classical linear mapping of *z* onto the IQ metric. Its benchmark record is the argument against it: reconstructing published battery conversions from their own structural inputs, GMEC-1 achieved a scale-balanced mean absolute error of 1.00 IQ points where the linear rule achieved 1.58 (WAIS-III development benchmark), and 1.25 against 3.13 on the Beta-III, whose upper tail the linear rule missed by 6.83 points on average—published batteries do curve their conversions in the tails, and by practically material amounts. On the JCCES the two routes agree within 0.3 points across the central range and separate only in the upper tail, by 2.0 points at a CAI sum of 45 and 9.0 at the attainable maximum. The second route is a direct percentile normalization of the summed standard scores—the conventional norming table. The study's norming simulations locate its weakness exactly where a conversion is hardest to estimate: the sampling error of a raw lookup table triples from the center of the scale to its tails, and at age-band sample sizes the table simply has no entry for a quarter to a third of the attainable range, whereas the formula's structural input—the subtest intercorrelations—is external and correctable.

Table 6.2 reprints the archived conversions at representative sums of standard scores.

Table 6.2

Archived conversions at representative sums of standard scores

Index	Sum	*z*	Converted score	95% interval
CAI	19	−1.40	79.4	[73.9, 84.9]
	30	0.00	100.0	[94.4, 105.6]
	45	+1.91	130.6	[124.0, 137.2]
	57	+3.43	160.5	[152.1, 168.8]
VAI	13	−1.24	81.4	[75.0, 87.9]
	20	0.00	100.0	[93.6, 106.4]
	30	+1.77	128.2	[120.8, 135.6]
	38	+3.19	156.3	[146.5, 166.0]

Note. Reprinted from Jouve (2026a, Table 5). Converted score = GMEC-1 value at the tabled sum; the interval is the delta-method 95% measurement interval around it, an uncertainty statement for an individual's score, not for the conversion. Full tables over the attainable ranges are archived with the study. CAI = Cognitive Acumen Index; VAI = Verbal Acumen Index.

Index Precision and Reporting Rules

On the standardization sample the composite reliabilities are .963 (CAI) and .953 (VAI) by stratified α, for classical standard errors of 2.88 and 3.25 IQ points (Cronbach, Schönemann, & McKie, 1965; Jouve, 2026a). Because the conversion is nonlinear, the archived intervals of Table 6.2 use the score-dependent delta-method error,

$$SEM_Q(S) \;=\; 15\,\sqrt{1-\alpha}\;G'(z),$$

which equals the classical error at the center and grows where the conversion stretches the metric: the tabled interval spans about ±6 points at a central CAI and about ±7 at a CAI of 130.6. That widening is honesty about a nonlinear scale, not lost precision.

Reports give each index as an integer with its 95% interval and a Wechsler-style qualitative band (Wechsler, 2008). Where the interval

straddles a band boundary, both bands are named; the interval, not the point, is the finding (AERA, APA, & NCME, 2014).

CHAPTER 7

Reliability and Precision

The internal-consistency record of the JCCES rests on two large samples analyzed years apart and, within the later one, on three independent computations. This chapter lays out that record and what it licenses.

The calibration study reported α = .96 for the full scale on its 1,079 protocols, with .81, .91, and .93 for VA, MP, and GK, .94 for the verbal scale, and a total-score standard error of 3.92 raw points (Jouve, 2023). On the 1,553-protocol standardization corpus the corresponding coefficients are .853, .914, .942, and .961 (Jouve, 2026b); the same three subtest values appear independently in the conversion study's reliability inputs (Jouve, 2026a) and in the parameter cross-validation study's scale summaries (Jouve & Sharma, 2026b). Between samples no coefficient moved beyond a few hundredths, which is the signature of coefficients determined by the item pool rather than by sampling accidents. Table 7.1 fixes the operative values.

Table 7.1

Internal consistency and index-level measurement error

Score	Coefficient	Value	*SEM*
Cognitive Acumen Index	stratified α	.963	2.88
Verbal Acumen Index	stratified α	.953	3.25
Verbal Analogies	α	.853	—
Math Problems	α	.914	—
General Knowledge	α	.942	—
Total scale (129 items)	α	.961	—

Note. Standardization sample, N = 1,553 (Jouve, 2026a; Jouve, 2026b). SEM is stated in IQ points for the indices, whose precision governs interpretation; subtest-level precision enters practice through the difference base rates of chapter 10. For dichotomous items α coincides with KR-20 (Cronbach, 1951).

Composite reliability is estimated by stratified α because an index built from heterogeneous subtests is exactly the stratified case: each subtest is a stratum, the between-stratum covariance is genuine signal, and the resulting coefficient properly exceeds every within-stratum α (Cronbach, Schönemann, & McKie, 1965). The practical content of Table 7.1 is the ±6-point 95% band that a CAI carries (Table 6.2 prints [94.4, 105.6] at the center) and the reminder that subtest scores, at .853 to .942, are dependable enough to build indices from but not precise enough to compare casually, which is why chapter 10 disciplines every subtest comparison with base rates. The ordering of the three coefficients follows length and homogeneity, GK being both the longest subtest and the most uniform in task demand. Score-level error away from the center follows the delta-method form of chapter 6, wider in the tails because the metric is stretched there; the two error statements agree wherever the conversion is effectively linear (Harvill, 1991).

CHAPTER 8

Internal Structure

The construct claim commits the JCCES to a particular anatomy: strong enough common variance across three content-distinct subtests to justify one composite, with interpretable specificity confined mainly to the quantitative facet. The published structural record tests that claim four ways, from the subtest level down to single items and outward against six external instruments.

Subtest Correlations

Table 8.1

Subtest intercorrelations and correlations with the CAI

Score	VA	MP	GK
Math Problems	.599		
General Knowledge	.774	.566	
Cognitive Acumen Index	.890	.841	.875

Note. N = 1,553 (Jouve, 2026b; Jouve, 2026h). *VA = Verbal Analogies; MP = Math Problems; GK = General Knowledge; CAI = Cognitive Acumen Index.*

The matrix already sketches the anatomy. The verbal pair correlates .774; MP joins each verbal subtest at .566 to .599, close enough to share a core, far enough to keep something of its own; and all three track the composite at .841 to .890 (Horn & Cattell, 1966).

The Factor Model

Confirmatory modeling on 11 item parcels of the standardization corpus formalizes the sketch (Jouve, 2026b). Before any model, the parcel eigenvalues announce a dominant dimension: 6.01 for the first, 54.6% of the variance, 5.14 times the second. Among the fitted models (Table 8.2), three correlated content factors beat one factor decisively ($\Delta\chi^2$ = 1,370.0, $\Delta df = 3$, $p < .001$), with latent correlations of .890 between the verbal factors and .694 and .609 for MP's pairings; a bifactor solution fits best, its GK specific factor collapsing to a boundary. The near-unity of the VA–GK latent correlation is worth noticing: it is why chapter 10 will show VA–GK differences to be the rarest kind.

Table 8.2

Confirmatory model fit, standardization sample

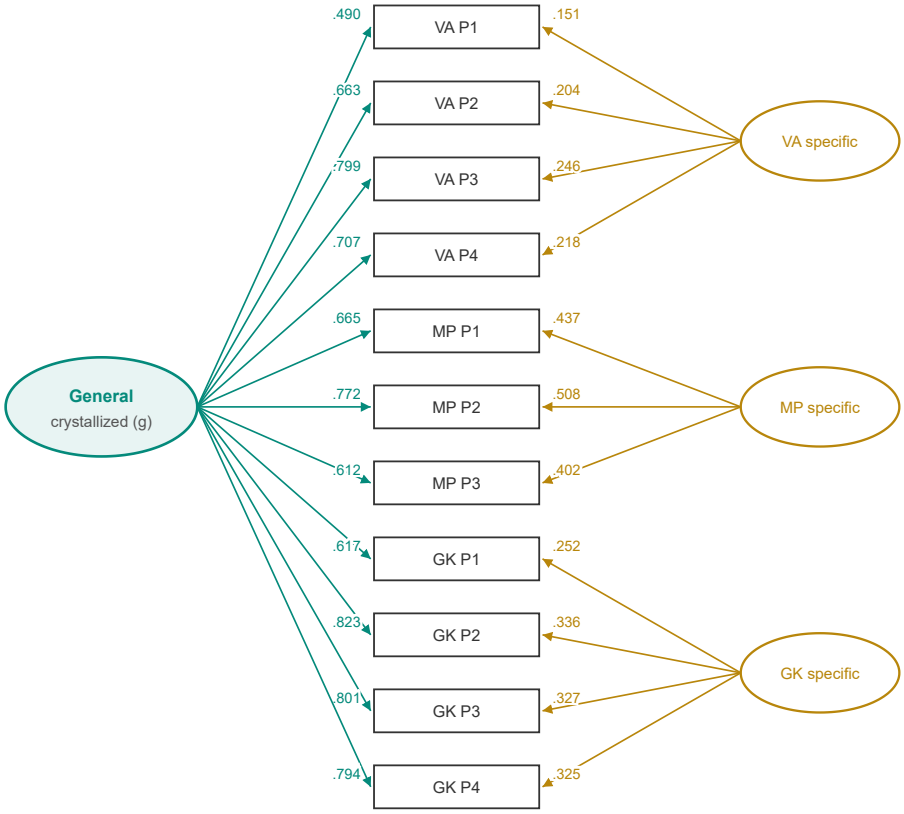
Model	χ^2	*df*	CFI	TLI	RMSEA	SRMR
Single factor	2,380.15	44	.786	.733	.185	.086
Three correlated factors	1,010.16	41	.911	.881	.123	.056
Higher-order *g*	1,010.45	41	.911	.881	.123	.056
Bifactor	538.87	33	.954	.923	.099	.037

Note. Maximum likelihood on 11 parcels, N = 1,553, all χ^2 $p < .001$ (Jouve, 2026b). The study's own simulation benchmark puts a correctly specified three-factor model near RMSEA = .063 at this design, the reference against which the absolute indices are read. CFI = comparative fit index; TLI = Tucker–Lewis index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual.

What settles interpretation is the variance accounting: ω_h = .879 and ω_t = .937, so the general dimension owns 93.7% of the reliable total-score variance and 82.6% of the common variance (ECV = .826), while the specific factors keep ω_s = .069 (VA), .260 (MP), and .128 (GK). Figure 8.1 draws the orthogonalized solution with its published standardized loadings: general loadings run .490 to .823, median .707, and every parcel loads higher on the general factor than on its specific one, with the specific loadings largest among the Math Problems parcels. An item-level replication on the tetrachoric matrix of all 129 items reproduces the ordering and the dominance (ECV = .822; .819 with six extreme items excluded). Form is three-factored; substance is general, with MP the one subtest whose residual is big enough to mean something (Jouve, 2026b).

Figure 8.1

Bifactor (Schmid–Leiman) solution for the 11 item parcels



Note. Standardized loadings from the Schmid–Leiman orthogonalization of the three-factor solution (Jouve, 2026b, Table 5); N = 1,553. Latent variables are drawn as ellipses and observed parcels as rectangles; general-factor paths are teal, specific-factor paths gold; residual terms are omitted for clarity. The decomposition yields ωh = .879, ωt = .937, and ECV = .826. VA = Verbal Analogies; MP = Math Problems; GK = General Knowledge.

The Geometry Against an External Battery

Where factoring counts variance, scaling draws the map. Scaled together with the four ACT sections from their 21 pairwise dissimilarities (n = 57), the seven scores need exactly two dimensions (Table 8.3), and the recovered plane (Figure 8.2) has a readable axis system: verbal-knowledge tasks stand opposite quantitative and science tasks, and the three JCCES subtests stand opposite the four ACT sections. Exact permutation tests make both directions dependable, the content effect at b = +.14 and the instrument effect at b = +.18, each at p = .0016, stable in direction across all 1,000 bootstrap resamples and intact after disattenuation (Jouve, 2016).

Table 8.3

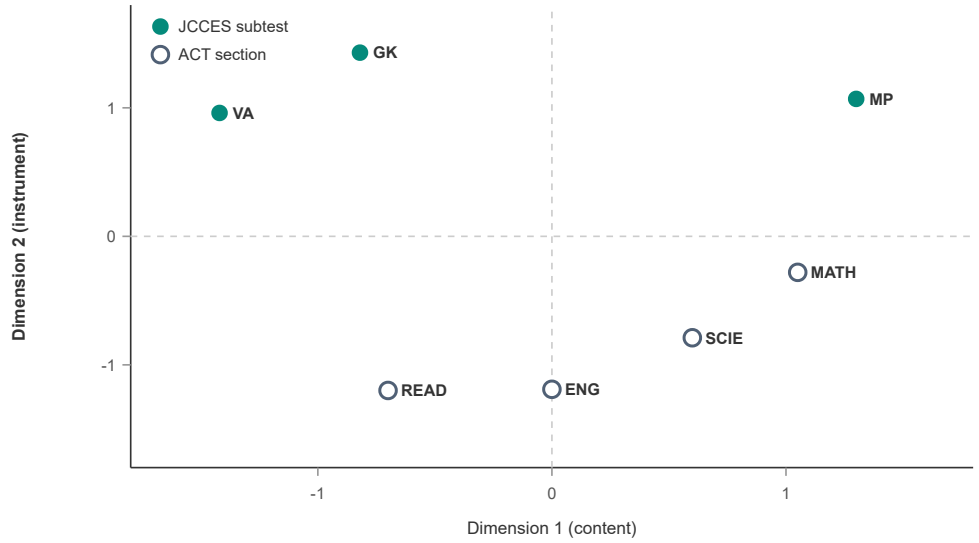
Scaling adequacy by dimensionality

Dimensions	Stress-1	RSQ
One	.339	.659
Two	.073	.960

Note. Ordinal ALSCAL, n = 57 (Jouve, 2016). A third dimension leaves residual stress of .008. RSQ = squared correlation of distances with disparities.

Figure 8.2

Joint configuration of the JCCES subtests and ACT sections



Note. Coordinates from the two-dimensional ordinal ALSCAL solution (Jouve, 2016, Table 2; n = 57). Filled circles are JCCES subtests, open circles ACT sections. VA = Verbal Analogies; MP = Math Problems; GK = General Knowledge; ENG, MATH, READ, SCIE = ACT English, Mathematics, Reading, Science. Verbal-knowledge tasks occupy the left of the plane and quantitative-science tasks the right; the JCCES subtests hold the upper half and the ACT sections the lower.

The same study's trait–method reading quantifies what the map shows: correlations between content-matched scores of different instruments average .59, MP with ACT Mathematics leading at .69, against .47 for pairs sharing neither content nor instrument, and the two composites correlate .69. Scores that carry similar labels on different batteries

measure kindred but not identical things, a caution every cross-battery comparison in chapter 9 inherits (Jouve, 2016).

Item-Level and Cross-Study Corroboration

Two further published angles agree. Pooling all 170 variance-bearing items of the JCCES and the What's Next number-sequence test, a NOHARM normal-ogive analysis finds one dominant dimension, Tanaka index .993 for the single-factor solution and only about .002 of residual improvement per added dimension, so even at item level, joined to a different instrument, the material is *g*-saturated (Jouve, 2025). And every criterion study of chapter 9 doubles as a structural probe: each ran a formal measurement-model contest against its criterion, and chapter 9 assembles those contests in Table 9.2 and Figure 9.1. Their joint verdict, fusion with crystallized criteria, organization by content in mixed company, separation from fluid ability, is the cross-battery form of the bifactor conclusion above.

CHAPTER 9

Criterion Evidence

Five published studies confront the indices with external instruments whose meanings are already settled: the clinical verbal indices of the RIAS and the Wechsler scales (Jouve, 2026d), the SAT in its 2005–2016 three-section edition (Jouve, 2026f), the pre-2011 GRE with its Analytical section set aside for want of cases (Jouve, 2026c), the AFQT read on an IQ metric whose published ceiling is the 99th percentile (Jouve, 2026e), and, as the designed counter-case, the nonverbal GAMA (Jouve, 2026g; Naglieri & Bardos, 1997). Between them the designs cover examiner-administered and group-administered criteria, verbal and figural content, and the educational, military, and clinical scoring traditions. Criterion instruments are documented in Reynolds and Kamphaus (2003) and Wechsler (2008) for the clinical batteries.

Because every sample is self-selected and able, each study publishes its coefficients in pairs, observed and range-corrected, with the

dispersion ratio *U* disclosed: univariate corrections after Thorndike (1949) for composites, multivariate Pearson–Lawley corrections for matrices. Restricted sampling suppresses a correlation without touching the regression it estimates, so the corrected value is each study's primary estimate and the observed value its floor; Table 9.1 prints both, so the reader can weigh the size of every adjustment. In the AFQT study the primary analysis is additionally ceiling-matched, both instruments capped at their common 99th-percentile maximum so neither is scored on headroom the other lacks; uncapping alters no coefficient by more than .02.

Table 9.1

Composite validities against the external criteria

Criterion	*n*	*U*	CAI *r*	CAI r_c	VAI *r*	VAI r_c
Crystallized and academic criteria						
RIAS Verbal Intelligence Index	119	1.16	.80	.84	.83	.87
WAIS-IV Verbal Comprehension Index	56	0.97	.83	.82	.86	.85
SAT composite	105	1.20–1.40	.74	.83	.65	.76
AFQT, IQ metric[a]	62	1.53	.70	.83	.69	.83
GRE composite	65	1.32–1.87	.58	.79	—	—
Fluid-figural criterion (discriminant)						
GAMA IQ	57	1.18	.58	.65	.47	.54

Note. r = observed, r_c = range-corrected; U = population-to-sample dispersion ratio of the criterion (a range where several sections are corrected jointly). Sources by row: Jouve (2026d) for RIAS and WAIS-IV; Jouve (2026f); Jouve (2026e); Jouve (2026c), whose verbal contrasts are subtest-level and appear in the text; Jouve (2026g). CAI = Cognitive Acumen Index; VAI = Verbal Acumen Index.

[a] Ceiling-matched analysis; full-width CAI values are .72 and .84.

Three readings organize the table. The first is the gradient: corrected validities of .76 to .87 wherever the criterion is crystallized or academic, dropping to .65 and .54 against the deliberately figural GAMA. The

latent versions of the same numbers, φ = .80 with the AFQT against φ = .47–.53 with the GAMA, repeat the gradient with reliability out of the way; the full set of measurement-model contests behind those values is assembled in Table 9.2 below. Convergence where the construct predicts it and separation where it forbids it is the whole discriminant argument, and the school-age Gc-achievement literature would have predicted the pattern (Floyd et al., 2003).

The second reading is content specificity inside the batteries. MP meets SAT Mathematics at .75 observed, .80 corrected (Jouve, 2026f); in the GRE design the content-matched cells run .62 to .74 corrected while the crossed cells sit at .04 to .12 (Jouve, 2026c); VA finds WAIS-IV Vocabulary at .82 and GK finds Information at .84 (Jouve, 2026d); and the AFQT, itself a blend, correlates .60 to .65 with all three subtests indistinctly (Jouve, 2026e). In the joint factor spaces the JCCES subtests carry loadings interchangeable with the criterion markers: .92 beside the RIAS indicators, .92 for VA and GK beside Vocabulary's .92 and Information's .91 in the WAIS-IV analysis, .90 to .92 corrected in the AFQT space (Jouve, 2026d; Jouve, 2026e). The GAMA study completes the picture from the far side: only MP bridges toward the fluid factor, at a cross-loading of .37, exactly where investment theory says the bridge should stand (Jouve, 2026g; Carroll, 1993).

The third reading concerns the norms themselves. Every convergent row of Table 9.1 checks the JCCES metric against an external norming program, and the AFQT study states the check in means as well as ranks: CAI minus AFQT averaged −0.44 points (d_z = −.05, nonsignificant), half of all pairs within 5 points, 81% within 10 (Jouve, 2026e). Two scales normed by different organizations on different populations land at the same level. The spread of those differences (*SD* = 8.77) is the equally important half of the sentence: agreement is a property of the calibrations, not a license to swap individual scores.

Latent-Variable Evidence

Each criterion study went beyond correlation to a formal measurement-model contest: candidate confirmatory models were fitted to the joint JCCES-and-criterion matrix, and the winning structure, with its latent

factor correlation φ, is the study's deepest finding. Table 9.2 assembles the published contests, and their joint reading is the strongest construct evidence the scale possesses, because the same three-subtest measurement model was confronted with five different criteria and the latent answer moved exactly as the construct claim requires.

Table 9.2

Measurement-model comparisons across the criterion studies

Study and model	χ^2	*df*	CFI	RMSEA	φ
JCCES + RIAS verbal subtests (*N* = 119; Jouve, 2026d)					
One crystallized factor (adopted)	12.03	5	.980	.109	—
Battery-split two factors	10.72	4	.981	.119	1.04[a]
JCCES + WAIS-IV verbal subtests (*N* = 56; Jouve, 2026d)					
One crystallized factor (adopted)	38.75	9	.889	.245	—
Battery-split two factors	38.28	8	.887	.262	1.02[a]
JCCES + SAT sections (*N* = 105; Jouve, 2026f)					
One common factor	74.2	9	.809	.264	—
Two factors, by battery	66.8	8	.827	.266	.87
By battery, quantitative doublet freed (adopted)	6.3	7	1.00	.000	.83
Two factors, by content	27.4	8	.943	.153	.68
JCCES + GRE sections (*N* = 65; Jouve, 2026c)					
One common factor	65.2	5	.605	.434	—
Two factors, by content trait (adopted)	16.9	4	.898	.224	.25
Two factors, by instrument	58.8	4	.644	.462	.76
JCCES + AFQT (*N* = 62; Jouve, 2026e)					
One crystallized factor (adopted)	1.91	2	1.00	0	.80[b]
JCCES + GAMA subtests (*N* = 57; Jouve, 2026g)					
One common factor	68.1	14	.679	.263	—

Study and model	χ^2	*df*	CFI	RMSEA	φ
Two factors, Gc–Gf	36.0	13	.830	.178	.53
Gc–Gf with MP cross-loading (adopted)	27.7	12	.870	.153	.49

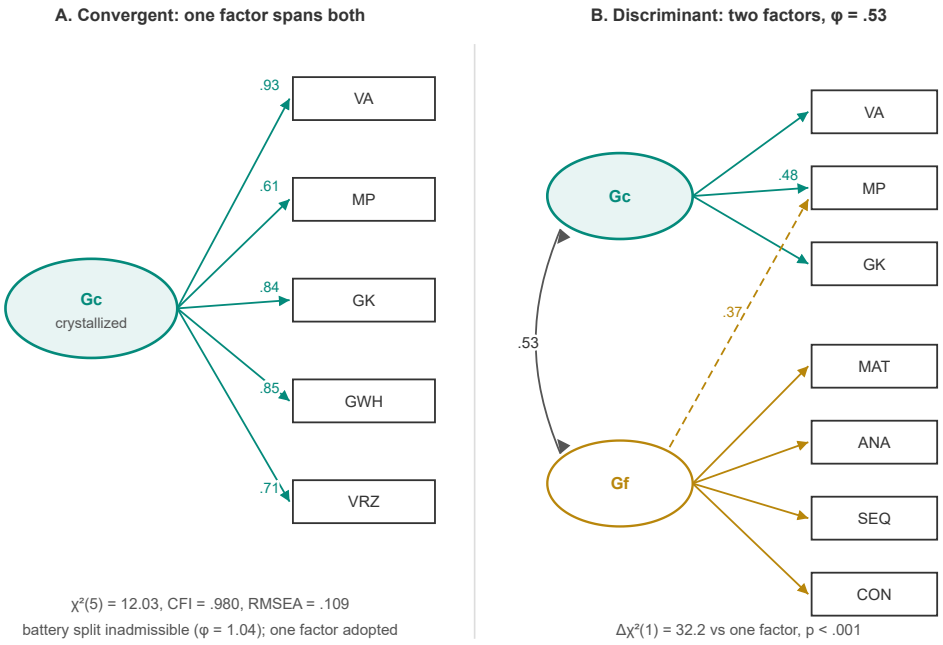
Note. Fit statistics as printed in each source study; the clinical and GAMA analyses are on range-corrected matrices, per their Methods. φ = correlation between the two latent factors of the model. CFI = comparative fit index; RMSEA = root mean square error of approximation; MP = Math Problems; Gc = crystallized ability; Gf = fluid ability.

[a] *Boundary value: the two-factor solution is inadmissible, so one factor is adopted.*

[b] *Disattenuated correlation from an equivalent reparameterization with the AFQT's composite reliability fixed; sensitivity range .79–.81.*

Figure 9.1

The two poles of the latent gradient: one shared factor with the RIAS, two correlated factors with the GAMA



Note. The two poles of the latent gradient of Table 9.2. Panel A: the adopted single-factor model of the JCCES subtests with the RIAS verbal subtests (N = 119), the largest and only examiner-administered criterion sample; standardized loadings as published in the study's CFA figure, adopted after the battery-split alternative proved inadmissible (Jouve, 2026d). Panel B: the adopted crystallized–fluid model with the GAMA on the range-corrected matrix; the dashed path is Math Problems' cross-loading, and the remaining standardized loadings are not reported in the source and are therefore unlabeled (Jouve,

2026g). Latent variables in ellipses, observed scores in rectangles; residuals omitted. Gc = crystallized; Gf = fluid; VA = Verbal Analogies; MP = Math Problems; GK = General Knowledge; GWH = RIAS Guess What; VRZ = RIAS Verbal Reasoning; MAT, ANA, SEQ, CON = GAMA Matching, Analogies, Sequences, Construction.

Reading down the φ column turns five studies into one experiment. Where the criterion is a clinical verbal index, the model cannot even hold the batteries apart: the two-factor correlation escapes to the boundary (1.04, 1.02) and collapses to a single crystallized factor spanning both instruments. Where the criterion is a broad academic or enlistment composite, one shared dimension again carries the data, with a latent correlation of .80 against the AFQT and .83 in the SAT model once its quantitative doublet is freed; the nested tests quantify the residual battery separation ($\Delta\chi^2(1) = 7.4$, $p = .007$ for the SAT, against 1.31 and 0.47, both nonsignificant, for the clinical batteries). Where content varies within the criterion, content wins: the GRE data prefer trait factors correlating .25 over instrument factors at .76, a $\Delta\chi^2(1)$ of 48.3 in favor of content organization. And where the criterion is deliberately fluid, the separation the construct demands appears on cue: a single factor is rejected against the GAMA ($\Delta\chi^2(1) = 32.2$, $p < .001$) and the factors settle at φ = .47–.53, with Math Problems the lone bridge (.48 on Gc, .37 on Gf). Figure 9.1 draws the two poles of that gradient. No single coefficient in Table 9.1 could establish this pattern; the model contests do.

The source studies attach their own cautions, which carry over verbatim: SAT, GRE, and AFQT scores were self-reported; the confirmatory samples run smaller than the conventional comfort line for latent modeling, with the WISC-III pilot ($n = 16$; Wechsler, 1991) too small to table here; and every corrected coefficient inherits the direct-selection model and the criterion's published dispersion, which is why the observed floors are printed beside them.

CHAPTER 10

Score Interpretation and Profile Analysis

The variance accounting of chapter 8 dictates the order of interpretation. With 93.7% of reliable total-score variance owned by the general dimension (Jouve, 2026b), a subtest score is mostly composite plus noise, and interpretation that starts at the bottom converts noise into narrative. Practice therefore runs top-down: CAI first, VAI when the question is verbal, subtest contrasts last and only when the base rates below certify them as unusual. Each reported score travels with its interval (chapter 7).

Choosing between the indices is a referral-question decision. The CAI is the default and the best-validated score (chapter 9). The VAI earns its place when verbal-crystallized functioning is itself the question, when quantitative performance is separately at issue, or when a certified CAI–VAI gap makes the aggregate misleading; it does not replace the CAI for general questions. Recurring configurations, verbal-over-quantitative, quantitative-over-verbal, an elevated or depressed knowledge subtest, take their conventional meanings from the interpretive literature (Kamphaus, 2001; Sattler, 2018), but no configuration means anything until its underlying difference clears the normative bar.

How Rare Is a Difference?

That bar is empirical. The discrepancy study tabulated, over the full standardization sample, how much of the population shows an absolute difference at or above each size, for the three subtest pairings and for the CAI–VAI contrast (Jouve, 2026h). Table 10.1 reprints the distributions, Figure 10.1 draws them, and Table 10.2 gives their summaries with the conventional 5% cutoffs. Mind the metrics: subtest rows are standard-score points (*SD* = 3), the CAI–VAI rows IQ points (*SD* = 15), so magnitudes do not compare across the two blocks.

Table 10.1

Percentage of the standardization sample at or above each absolute difference

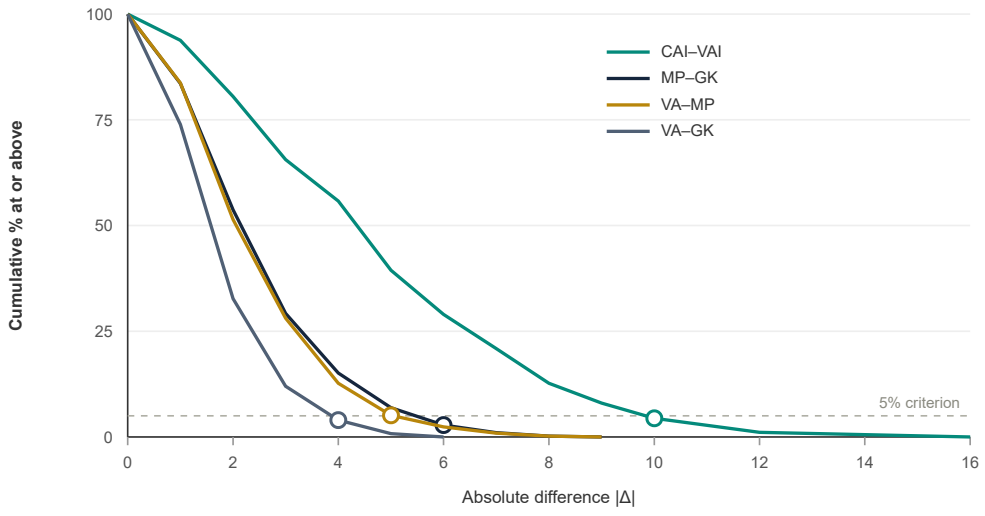
|Δ|	CAI–VAI	VA–MP	VA–GK	MP–GK
≥ 1	93.8	83.7	73.9	83.6

|Δ|	CAI–VAI	VA–MP	VA–GK	MP–GK
≥ 2	80.5	51.4	32.7	53.7
≥ 3	65.6	28.1	12.0	29.2
≥ 4	55.8	12.7	4.0	15.1
≥ 5	39.4	5.1	0.8	6.9
≥ 6	29.0	2.4	0.0	2.8
≥ 7	20.9	0.9	—	1.0
≥ 8	12.7	0.2	—	0.2
≥ 9	8.0	0.0	—	0.0
≥ 10	4.4	—	—	—
≥ 12	1.1	—	—	—
≥ 16	0.0	—	—	—

Note. N = 1,553 (Jouve, 2026h). Differences are absolute and carry no direction; subtest columns are on the standard-score metric, the CAI–VAI column on the IQ metric. CAI = Cognitive Acumen Index; VAI = Verbal Acumen Index; VA = Verbal Analogies; MP = Math Problems; GK = General Knowledge.

Figure 10.1

How common each size of score difference is, with the .05 cutoffs



Note. Curves plot the cumulative percentages of Table 10.1 (N = 1,553; Jouve, 2026h); open circles sit at the .05 cutoffs of Table 10.2. Subtest comparisons are

on the standard-score metric, the CAI–VAI comparison on the IQ metric. Only where a curve has descended to the dashed criterion line is a difference unusual. CAI = Cognitive Acumen Index; VAI = Verbal Acumen Index; VA = Verbal Analogies; MP = Math Problems; GK = General Knowledge.

Table 10.2

Difference distributions and 5% cutoffs

Comparison	*M*	*SD*	.05 cutoff
CAI–VAI	4.14	2.81	≥ 10
VA–MP	1.85	1.47	≥ 5
VA–GK	1.23	1.08	≥ 4
MP–GK	1.93	1.54	≥ 6

Note. N = 1,553 (Jouve, 2026h). The cutoff is the smallest absolute difference shown by no more than 5% of the sample; a difference beyond it is unusual, which is a statement about rarity, not about cause. Abbreviations as in Table 10.1.

The distributions repay a close look. VA and GK travel together, 26.1% of the sample showing no difference at all and only 4.0% reaching 4 points, so even moderate VA–GK separations are genuinely uncommon, the profile-level echo of the .890 latent correlation of chapter 8. Everything involving MP scatters more, echoing MP's ω_s = .260, and its contrasts certify only at 5 to 6 points. The CAI–VAI contrast, which isolates MP by construction, averages 4.14 IQ points and certifies at 10, a size reached by 4.4% of the population. Concretely: a 7-point CAI–VAI gap is matched or exceeded by one examinee in five and warrants silence, while a 4-point VA–GK gap sits in the top 4% and warrants a sentence. Rarity, not size, is what licenses comment.

For users who prefer a reliability-based check, the discrepancy study also documents the standard-error-of-difference test; the two criteria answer different questions, rarity versus measurement dependability, and a report says which one it applied (Jouve, 2026h). Whatever the arithmetic, educational and language history frame every profile

reading: crystallized scores are records of opportunity as well as ability (Sattler, 2018).

CHAPTER 11

Protocol Review

A score is only as interpretable as the response pattern behind it. Every JCCES protocol therefore passes a three-part review before interpretation, built from the calibrated item parameters of chapter 4 and computed over administered items only: does the pattern fit the model of the person (person fit), does success respect the difficulty ordering (monotonicity), and is the ability estimate sharp where this examinee stands (conditional precision)? The screens flag; they do not explain. Attribution of cause stays outside their reach (Drasgow et al., 1985; Embretson & Reise, 2000; AERA, APA, & NCME, 2014).

The three ask genuinely different questions. An examinee can reach an ordinary total through an extraordinary route, missing easy items while solving hard ones; the fit statistics notice the route. Monotonicity asks the same question without leaning on the parametric model, simply counting order violations. Conditional precision asks not whether the pattern is credible but whether the pool had enough information at this examinee's level to pin the estimate down. Because review is universal rather than triggered, its outcome is a routine line in every report, never an insinuation.

Person Fit

With expected success $p_i = \{1 + \exp[-a_i(\theta - b_i)]\}^{-1}$ and variance $v_i = p_i(1 - p_i)$ for item i, the review uses the weighted and unweighted mean squares with a standardized residual sum (Wright & Masters, 1982; Linacre, 2002; Meijer & Sijtsma, 2001):

$$\text{Infit MSQ} = \frac{\sum_i v_i\, z_i^2}{\sum_i v_i}, \quad \text{Outfit MSQ} = \frac{1}{k}\sum_i z_i^2, \quad z_i = \frac{X_i - p_i}{\sqrt{v_i}}, \quad Z_{\text{sum}} = \frac{\sum_i (}{\sqrt{}}$$

Monotonicity

With items ordered by difficulty, the reversal rate is the share of item pairs answered against that order (Guttman, 1950; Sijtsma & Molenaar, 2002):

$$\text{Reversal Rate} = \frac{\#\{(i,j) : b_i < b_j,\ X_i{=}0,\ X_j{=}1\}}{\#\{(i,j) : b_i < b_j\}}.$$

Conditional Precision

Precision at the examinee's level is the reciprocal square root of the test information there (Embretson & Reise, 2000; Baker & Kim, 2004):

$$SE(\theta) = I(\theta)^{-1/2}, \qquad I(\theta) = \sum_i a_i^2\, p_i(1 - p_i).$$

Combining the Indicators

Table 11.1 states the operational bands and the action each pattern triggers. Mean squares under 1.0 mark unusual predictability, which is benign and triggers nothing.

Table 11.1

Review outcome by indicator band

Outcome	Infit/Outfit MSQ	$\lvert Z_{sum} \rvert$	Reversal rate	SE(θ)	Action
Consistent & precise	≤ 1.3	< 2	< 10%	≤ 0.30	Interpret as obtained
Review	1.3–1.5	[2, 3)	10–20%	0.30–0.50	Add narrative qualifiers; consult the administration record
Caution	> 1.5	≥ 3	> 20%	> 0.50	Interpret with reservation; corroborate or retest

Note. Operational policy bands of the testing platform; the fit and reversal conventions are heuristic, not normative (Guttman, 1950; Sijtsma & Molenaar, 2002). The first outcome requires every criterion met; the others follow the single worst indicator. MSQ = mean square; Z_{sum} = standardized residual sum; SE(θ) = conditional standard error on the latent metric, smaller meaning sharper (Embretson & Reise, 2000).

The review's outcome is stated per score, in a fixed form:

> "Protocol review found all indicators within their expected bands (Table 11.1): fit statistics and reversal rates unremarkable, conditional precision adequate, no rapid-responding flags. The review supports interpretation of the scores as obtained."

CHAPTER 12

Uses, Populations, and Boundaries of the Evidence

What the JCCES may responsibly be used for follows from what chapters 7 through 9 show it measures: a reliable, criterion-anchored index of crystallized ability. Uses that lean on that construct, educational assessment and placement, identification of verbal giftedness, evaluation of academic attainment, knowledge-oriented occupational assessment, research on Gc, and supplementary testing inside a fuller battery, stand on published corrected validities of .76 to .87 against clinical and academic criteria. Uses that would require fluid reasoning, processing speed, or nonverbal ability find the opposite answer in the

same record, the GAMA separation of chapters 8 and 9, and call for a different instrument alongside.

Boundaries

- Administration is in English only; developing English proficiency depresses scores for language reasons
- Responses are typed; motor or entry constraints can masquerade as ability
- No single instrument grounds a diagnosis (AERA, APA, & NCME, 2014)
- Knowledge items are cultural products; opportunity to learn shapes performance on them
- Fluid reasoning and processing speed are outside the measured construct
- The validity samples are self-selected and able; the published studies pair every corrected coefficient with its observed floor for exactly that reason (chapter 9)

Populations

The best-supported use is with English-proficient adults and older adolescents of at least middle-school education; the calibration corpus itself spanned ages 8 to 72 with a mean of 30.14 (Jouve, 2023), but the criterion record rests on adult samples. For examinees whose schooling was interrupted, conducted in another language, or marked by language-based learning difficulty, obtained scores can read low for reasons the construct does not intend; the interpretive machinery of chapter 10 and the review of chapter 11 assist that judgment without replacing it.

The Second Edition

A successor instrument, the JCCES-II, is in development: six subtests and 260 items spanning linguistic analogy and sequencing, algebraic and geometric reasoning, and humanities and quantitative-scientific

information, with three Verbal Analogies anchor items retained from the first-edition bank. Its published pilot (N = 62) reports subtest KR-20 of .881 to .949, a 260-item total at .978, a single factor retained by parallel analysis but reliable structure beyond g on confirmatory comparison (general-factor ECV = .69, ω_h = .82), a correlation of .91 with the JCCES-I CAI in the paired subsample (n = 25), and the provisional equating $CAI_{eq} = 66.8 + 0.369 \times S$ (Jouve, 2024). Those anchors calibrate the linkage, not population norms; until a standardization sample exists, the first edition documented here remains the operational instrument.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). Lawrence Erlbaum Associates. https://doi.org/10.1201/9781482276725

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press. https://doi.org/10.1017/CBO9780511571312

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297-334. https://doi.org/10.1007/BF02310555

Cronbach, L. J., Schönemann, P., & McKie, D. (1965). Alpha coefficients for stratified-parallel tests. *Educational and Psychological Measurement, 25*(2), 291–312. https://doi.org/10.1177/001316446502500201

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *Applied Psychological Measurement, 9*(4), 297–308. https://doi.org/10.1111/j.2044-8317.1985.tb00817.x

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.

Floyd, R. G., Evans, J. J., & McGrew, K. S. (2003). Relations between measures of Cattell-Horn-Carroll (CHC) cognitive abilities and

mathematics achievement across the school-age years. *Psychology in the Schools*, 40(2), 155-171. https://doi.org/10.1002/pits.10083

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer (Ed.), *Measurement and prediction* (pp. 60–90). Princeton University Press.

Harvill, L. M. (1991). Standard error of measurement. *Educational Measurement: Issues and Practice, 10*(2), 33–41. https://doi.org/10.1111/j.1745-3992.1991.tb00195.x

Horn, J. L., & Blankson, N. (2005). Foundations for better understanding of cognitive abilities. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 41–68). Guilford Press.

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology*, 57(5), 253-270. https://doi.org/10.1037/h0023816

Jouve, X. (2016). Content and instrument organization of JCCES and ACT subtests: A multidimensional scaling analysis. *Cogn-IQ Research Papers*. https://pubscience.org/ps-1mADL-b45fe5-Sjy6

Jouve, X. (2023). Evaluating the Jouve Cerebrals Crystallized Educational Scale (JCCES): Reliability, internal consistency, and alternative answer recognition. *Cogn-IQ Research Papers*. https://pubscience.org/ps-1mSR3-32426c-iYHT

Jouve, X. (2024). Jouve Cerebrals Crystallized Educational Scale II: Pilot calibration and equating to the JCCES-I Cognitive Acumen Index. *Cogn-IQ Research Papers*. https://pubscience.org/ps-1mVDO-9b28f7-uQHF

Jouve, X. (2025). Numbers as symbols: An item-level multidimensional scaling of the What's Next number-sequence test and the JCCES. *Cogn-IQ Research Papers*. https://pubscience.org/ps-1mYgv-e7bac6-1l7r

Jouve, X. (2026a). Beyond Tellegen–Briggs Formula 4: Development, external benchmarking, and prospective application of a generic monotone composite conversion. *Cogn-IQ Research Papers*. https://pubscience.org/ps-1n0JI-d36170-9ZCX

Jouve, X. (2026b). Confirmatory factor analysis of the Jouve Cerebrals Crystallized Educational Scale: Evidence for a three-factor structure with a dominant general factor. *Cogn-IQ Research Papers*. https://pubscience.org/ps-1n0I3-799113-Lpri

Jouve, X. (2026c). Convergent and discriminant validity of the JCCES with the Graduate Record Examinations: A multitrait–multimethod analysis. *Cogn-IQ Research Papers*. https://pubscience.org/ps-1n0GJ-29c263-HhFq

Jouve, X. (2026d). Convergent validity of the JCCES with individually administered intelligence batteries: The Reynolds (RIAS) and Wechsler (WAIS-IV, WISC-III) verbal indices. *Cogn-IQ Research Papers*. https://pubscience.org/ps-1n0Go-ead7e9-8PRA

Jouve, X. (2026e). Convergent validity of the JCCES with the AFQT: A crystallized battery and the military enlistment composite. *Cogn-IQ Research Papers*. https://pubscience.org/ps-1n0JL-42bd17-9PZD

Jouve, X. (2026f). Convergent validity of the JCCES with the SAT: Bivariate, general-factor, and latent-variable evidence. *Cogn-IQ Research Papers*. https://pubscience.org/ps-1n0Gv-3d028d-ZVOh

Jouve, X. (2026g). Discriminant validity of the JCCES against the GAMA: Evidence for a crystallized–fluid (Gc–Gf) separation. *Cogn-IQ Research Papers*. https://pubscience.org/ps-1n0GC-134d02-iTaY

Jouve, X. (2026h). Subtest score discrepancy analysis: Base rates of VA–MP–GK profile scatter in the JCCES. *Cogn-IQ Research Papers*. https://pubscience.org/ps-1n0D8-2c335f-8pZQ

Jouve, X., Wai, J., Watterson, J., Woodley, M., Faverio, M. C., Lindgren, K., Whitley, J., & Freitas, N. (2023). *Jouve-Cerebrals Crystallized-Educational Scale (JCCES) – Revised*. Cogn-IQ Cognitive Assessments. https://pubscience.org/ps-1mVDO-e10f3a-Tc1I

Jouve, X., & Sharma, P. (2026a). Efficient answer keys for open-ended items: Coverage-optimal construction and the limits of normalization. *Cogn-IQ Research Papers*. https://pubscience.org/ps-1n0HU-6d0458-zFTN

Jouve, X., & Sharma, P. (2026b). The invariant hierarchy of an exchangeable response matrix: Self-calibrating group-theoretic estimation of item parameters. *Cogn-IQ Research Papers*. https://pubscience.org/ps-1n0HU-627ecf-UCzT

Jouve, X., & Sharma, P. (2026c). Option characteristic curves for open-ended items: Data-driven answer selection and scoring. *Cogn-IQ Research Papers*. https://pubscience.org/ps-1n0HU-95acf3-Pmli

Kamphaus, R. W. (2001). *Clinical assessment of child and adolescent intelligence* (2nd ed.). Allyn & Bacon.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25*(2), 107–135. https://doi.org/10.1177/01466210122031957

Naglieri, J. A., & Bardos, A. N. (1997). *General Ability Measure for Adults*. National Computer Systems.

Reynolds, C. R., & Kamphaus, R. W. (2003). *Reynolds Intellectual Assessment Scales: Professional manual*. PAR.

Sattler, J. M. (2018). *Assessment of children: Cognitive foundations and applications* (6th ed.). Jerome M. Sattler Publisher.

Schneider, W. J., & McGrew, K. S. (2018). The Cattell-Horn-Carroll theory of cognitive abilities. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (4th ed., pp. 73–163). Guilford Press.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Sage. https://doi.org/10.4135/9781412984676

Tellegen, A., & Briggs, P. F. (1967). Old wine in new skins: Grouping Wechsler subtests into new scales. *Journal of Consulting Psychology, 31*(5), 499–506. https://doi.org/10.1037/h0024963

Thorndike, R. L. (1949). *Personnel selection: Test and measurement techniques*. Wiley.

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children–Third Edition (WISC-III)*. The Psychological Corporation.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale–Fourth Edition (WAIS-IV)*. Pearson.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163–183. https://doi.org/10.1207/s15324818ame1802_2

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. MESA Press.